Universidad de San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Ingeniería Mecánica Industrial

CONSTRUCCIÓN DE UN MODELO DE REGRESION LOGÍSTICA PARA ESTIMAR LA PROBABILIDAD DE RETENCIÓN DE CLIENTES PROFESIONALES, EN UNA EMPRESA MINORISTA DE MATERIALES DE CONSTRUCCIÓN DE GUATEMALA

Pedro Pablo Morales Ortíz
Asesorado por Ing. José Rolando Chávez Salazar

Guatemala, febrero de 2023

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

CONSTRUCCIÓN DE UN MODELO DE REGRESION LOGÍSTICA PARA ESTIMAR LA PROBABILIDAD DE RETENCIÓN DE CLIENTES PROFESIONALES, EN UNA EMPRESA MINORISTA DE MATERIALES DE CONSTRUCCIÓN DE GUATEMALA

TRABAJO DE GRADUACIÓN

PRESENTADO A LA JUNTA DIRECTIVA DE LA
FACULTAD DE INGENIERÍA
POR

PEDRO PABLO MORALES ORTÍZ
ASESORADO POR ING. JOSÉ ROLANDO CHÁVEZ SALAZAR

AL CONFERÍRSELE EL TÍTULO DE

INGENIERO MECÁNICO INDUSTRIAL

GUATEMALA, FEBRERO DE 2023

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

NÓMINA DE JUNTA DIRECTIVA

DECANO	Inga. Aurelia Anabela Cordova Estrada
VOCAL I	Ing. José Francisco Gómez Rivera
VOCAL II	Ing. Mario Renato Escobedo Martínez
VOCAL III	Ing. José Milton de León Bran
VOCAL IV	Br. Christian Moisés de la Cruz Leal
VOCAL V	Br. Kevin Vladimir Armando Cruz Lorente
SECRETARIA	Ing. Hugo Humberto Rivera Pérez

TRIBUNAL QUE PRACTICÓ EL EXAMEN GENERAL PRIVADO

DECANO	Ing. Pedro Antonio Aguilar Polanco
EXAMINADORA	Inga. Alba Maritza Guerrero Spinola
EXAMINADOR	Ing. Erwin Danilo Gonzales Trejo
EXAMINADOR	Ing. Edgar Darío Álvarez Coti
SECRETARIA	Inga. Lesbia Magalí Herrera López

HONORABLE TRIBUNAL EXAMINADOR

En cumplimiento con los preceptos que establece la ley de la Universidad de San Carlos de Guatemala, presento a su consideración mi trabajo de graduación titulado:

CONSTRUCCIÓN DE UN MODELO DE REGRESION LOGÍSTICA PARA ESTIMAR LA PROBABILIDAD DE RETENCIÓN DE CLIENTES PROFESIONALES, EN UNA EMPRESA MINORISTA DE MATERIALES DE CONSTRUCCIÓN DE GUATEMALA

Tema que me fuera asignado por la dirección de la Escuela de Ingeniería Mecánica Industrial, con fecha 10 de noviembre de 2022.

Pedro Pablo Morales Ortíz

ESCUELA DE ESTUDIOS DE
POSTGRADO
FACULTAD DE INGENIERÍA

EEPFI-PP-1787-2022
Guatemala, 10 de noviembre de 2022

Director
César Ernesto Urquizú Rodas
Escuela Ingenieria Mecanica Industrial
Presente.

Estimado Ing. Urquizú

Reciba un cordial saludo de la Escuela de Estudios de Postgrado de la Facultad de Ingeniería.

El propósito de la presente es para informarle que se ha revisado y aprobado el Diseño de Investigación titulado: **CONSTRUCCIÓN DE UN MODELO DE REGRESION LOGÍSTICA PARA ESTIMAR LA PROBABILIDAD DE RETENCIÓN DE CIENTES PROFESIONALES EN UNA EMPRESA MINORISTA DE MATERIALES DE CONSTRUCCIÓN EN GUATEMALA**, el cual se enmarca en la línea de investigación: **Todas las áreas - Análisis de datos categóricos**, presentado por el estudiante **Pedro Pablo Morales Ortiz** carné número **201403531**, quien optó por la modalidad del "PROCESO DE GRADUACIÓN DE LOS ESTUDIANTES DE LA FACULTAD DE INGENIERÍA OPCIÓN ESTUDIOS DE POSTGRADO". Previo a culminar sus estudios en la Maestría en ARTES en Estadistica Aplicada.

Y habiendo cumplido y aprobado con los requisitos establecidos en el normativo de este Proceso de Graduación en el Punto 6.2, aprobado por la Junta Directiva de la Facultad de Ingeniería en el Punto Décimo, Inciso 10.2 del Acta 28-2011 de fecha 19 de septiembre de 2011, firmo y sello la presente para el trámite correspondiente de graduación de Pregrado.

Atentamente,

"Id y Enseñad a Todos"

Ing. José Rolando Chávez Salazar
Ingeniero Industrial
Colegiado No. 4,317

Mtro. José Rolando Chávez Salazar
Asesor(a)

Mtro. Edwin Adalberto Bracamonte Orozco
Coordinador(a) de Maestría

Mtro. Edgar Darío Alvaréz Cotí
Director
Escuela de Estudios de Postgrado
Facultad de Ingeniería

UNIVERSIDAD DE SAN CARLOS
DE GUATEMALA

FACULTAD DE INGENIERÍA

EEP-EIMI-1441-2022

El Director de la Escuela Ingenieria Mecanica Industrial de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer el dictamen del Asesor, el visto bueno del Coordinador y Director de la Escuela de Estudios de Postgrado, del Diseño de Investigación en la modalidad Estudios de Pregrado y Postgrado titulado: **CONSTRUCCIÓN DE UN MODELO DE REGRESION LOGÍSTICA PARA ESTIMAR LA PROBABILIDAD DE RETENCIÓN DE CIENTES PROFESIONALES EN UNA EMPRESA MINORISTA DE MATERIALES DE CONSTRUCCIÓN EN GUATEMALA**, presentado por el estudiante universitario **Pedro Pablo Morales Ortiz**, procedo con el Aval del mismo, ya que cumple con los requisitos normados por la Facultad de Ingeniería en esta modalidad.

ID Y ENSEÑAD A TODOS

UNIVERSIDAD DE SAN CARLOS
DIRECCION
Escuela de Ingeniería Mecánica Industrial
FACULTAD DE INGENIERIA

Ing. César Ernesto Urquizú Rodas
Director
Escuela Ingenieria Mecanica Industrial

Guatemala, noviembre de 2022

Decanato
Facultad de Ingeniería
24189101- 24189102
secretariadecanato@ingenieria.usac.edu.gt

LNG.DECANATO.OI.205.2023

La Decana de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer la aprobación por parte del Director de la Escuela de Ingeniería Mecánica Industrial, al Trabajo de Graduación titulado: **CONSTRUCCIÓN DE UN MODELO DE REGRESION LOGÍSTICA PARA ESTIMAR LA PROBABILIDAD DE RETENCIÓN DE CLIENTES PROFESIONALES, EN UNA EMPRESA MINORISTA DE MATERIALES DE CONSTRUCCIÓN DE GUATEMALA,** presentado por: **Pedro Pablo Morales Ortíz,** después de haber culminado las revisiones previas bajo la responsabilidad de las instancias correspondientes, autoriza la impresión del mismo.

IMPRÍMASE:

Inga. Aureiia Anabela Cordova Estrada
Decana

Guatemala, febrero de 2023

AACE/gaoc

ACTO QUE DEDICO A:

Mis padres	Por guiarme siempre por el buen camino, buscando mi superación personal
Mis abuelos	Por su constante apoyo en mi proceso de aprendizaje y motivación para continuar en los momentos más difíciles.
Helena Reyes	Por ser mi apoyo incondicional y mi fuente de motivación
Mis amigos	Por brindarme siempre su ayuda y conocimientos durante nuestra etapa de estudios.

AGRADECIMIENTOS A:

Universidad de San Carlos de Guatemala	Por brindarme una educación de calidad y un segundo hogar.
Mi Familia	Por brindar un entorno de confianza para mi desarrollo académico.
Ing. Rolando Chávez	Por su apertura a brindar asesoría en el proceso y los buenos consejos.
Ing. Carlos Beltetón	Por su apoyo en el desarrollo de este proyecto en su unidad de negocio.
Ing. William Fagiani	Por su incontable apoyo para la elección y calibración de los modelos más adecuados.
Mis catedráticos	Por brindarme su incontable conocimiento y consejos durante mi etapa estudiantil.

ÍNDICE GENERAL

ÍNDICE DE ILUSTRACIONES

FIGURAS

TABLAS

LISTA DE SÍMBOLOS

Símbolo	Significado
F	Frecuencia de compra
H1	Hipótesis alterna
H0	Hipótesis nula
M	Monto promedio
%	Porcentaje
R	Reciencia

GLOSARIO

Centroide Es la ubicación real o imaginaria que representa el centro de un grupo de datos. También puede considerarse como el centro geométrico en un espacio de k variables o dimensiones.

Clientes Profesionales Son los clientes que, por su modelo de operación comercial, proveen servicios a otros clientes finales. Se encuentran identificados de esta forma en la base de datos de la empresa.

Clúster Son agrupaciones de datos o registros de datos que comparten características o están relacionadas entre sí.

ERP Se refiere a un *software,* Enterprise Resource Planning*,* que se traduce como sistema de planificación de recursos empresariales. Es un sistema que ayuda a administrar los procesos contables y comerciales de una empresa, registrando sus procesos de ventas, cadena de suministro, operaciones, recursos humanos, entre otros.

Machine learning	Es una rama de la inteligencia artificial que, a través de diferentes algoritmos, da la capacidad a sistemas de información para identificar patrones en datos masivos y elaborar predicciones.
Minería de datos	Es la exploración y análisis de datos, automático y semiautomático, que analiza grandes cantidades de información para descubrir patrones o reglas que sean significativas.
Multicolinealidad	Es la relación de dependencia lineal entre dos o más variables independientes en un proceso de regresión.
Quintiles	Son cuatro valores que dividen un conjunto de datos ordenados en cinco segmentos del mismo tamaño.
Residuo estandarizado	Es la transformación que se obtiene al dividir un valor residual de un modelo dentro de su desviación estándar estimada.
Variable de intervalo	Es una variable de tipo cuantitativo, donde los intervalos entre sus clases son iguales; sin embargo, el cero no implica el valor nulo de un atributo.
Variable de razón	Variable de tipo cuantitativo donde el cero sí indica la ausencia total de un atributo, por lo que sí es posible hacer razones en la medición.

Variable dicotómica	Es un tipo de variable cualitativa que solo puede tomar dos valores que son mutuamente excluyentes, denotando la ausencia o presencia de una característica.
Variables ordinales	Son variables cualitativas donde cada clase posee una misma relación posicional con la siguiente, por lo que muestra situaciones escalonadas.

RESUMEN

La empresa minorista de materiales de construcción cuenta con una amplia cartera de clientes considerados profesionales. Por el tipo de modelo de negocio de la empresa, estos clientes pueden representar una fuerte cantidad de negocios futuros. Por esta razón, la empresa busca el desarrollo de lealtad sobre estos clientes.

El desarrollo de la lealtad en los clientes requiere conocer a profundidad la forma en que se comportan los clientes, y si estos están en riesgo de perderse. Durante la fase de formulación del problema, se evidenció la poca información predictiva sobre los clientes, lo cual imposibilita hacer seguimiento de forma estructurada.

Para solucionar este vacío de información, se propuso la creación de un modelo de regresión logística, utilizando la información transaccional de los clientes y sus dimensiones de clasificación. Con ello, se contará con un modelo adecuadamente calibrado, que indique la probabilidad de retención de cada cliente profesional.

El procedimiento para la modelización matemática está enfocado en brindar un modelo de datos con alta exactitud, utilizando únicamente las variables que tengan una correlación significativa con la variable de respuesta.

OBJETIVOS

General

Construir un modelo de regresión logística para estimar la probabilidad de retención de clientes profesionales, en una empresa minorista de materiales de construcción en Guatemala.

Específicos

1. Agrupar a los clientes profesionales en segmentos similares, basado en las variables de reciencia, frecuencia y monto de compras, aplicando métodos de simulación por K Medias.

2. Identificar las variables de clientes que interfieren en la pérdida o retención de clientes, usando pruebas de independencia y pruebas de correlación.

3. Relacionar las variables cuantitativas y cualitativas de los clientes profesionales, para construir un modelo de regresión logística que permita cuantificar las probabilidades de pérdida y compra de cada cliente profesional.

INTRODUCCIÓN

El presente estudio consiste en la sistematización del proceso de análisis y segmentación de clientes profesionales, mediante el cálculo de la probabilidad de retención, para su posterior tratamiento en la empresa minorista de materiales de construcción en Guatemala.

En un análisis preliminar, se determinó que la empresa no tiene la capacidad de desarrollar técnicas predictivas sobre el comportamiento de los clientes, y tampoco se ha estudiado con profundidad qué variables de segmentación de estos tienen una correlación significativa con la compra o pérdida de los clientes.

La importancia del estudio radica en la necesidad de la empresa de generar acciones comerciales para su crecimiento sostenido. Contar con este tipo de información permite diseñar estrategias de personalización masiva, que son de gran apoyo a la fidelización de cartera de clientes y a la atención de las necesidades de servicio en aquellos que lo requieran, incrementando la satisfacción del cliente y la rentabilidad de la empresa.

La metodología de la investigación es de enfoque cuantitativo, con diseño no experimental u observacional y un alcance descriptivo correlacional.

El resultado que se espera con el presente estudio es que la empresa tenga a su disposición una estimación puntual automatizada, mediante un modelo de regresión logística, de la probabilidad de retención de cada uno de

sus clientes profesionales, de forma que puedan segmentarlos e identificarlos rápidamente en sus sistemas de información.

El desarrollo del estudio estará compuesto por seis fases. En la primera fase se hará una recopilación bibliográfica, que será de utilidad para fundamentar los temas que se analizarán en esta investigación. En la segunda se realizarán tareas de minería de datos para extraer las variables de segmentación y de comportamiento de los clientes, utilizando documentos transaccionales del 2021 al 2022. La tercera fase del estudio comprende un análisis de correlación, para inferir si cada variable tiene un efecto significativo en la retención o pérdida de los clientes profesionales. La cuarta fase tomará las variables que durante el análisis de correlación probaron ser significantes en su efecto sobre la variable de respuesta, para construir un modelo de regresión logística binomial. Además, se evaluará la multicolinealidad de las variables elegidas para iterar nuevamente, consiguiendo múltiples modelos que se ajusten a los datos presentados. La quinta fase del estudio implica analizar los resultados para elegir el modelo de regresión con mejor ajuste. La última fase del estudio consiste en presentar los resultados obtenidos de la investigación, así como la redacción de conclusiones, recomendaciones e informe final.

La factibilidad del estudio es la adecuada, ya que se cuenta con todos los recursos tecnológicos, financieros y de información necesarios para su ejecución.

El informe final estará constituido por los siguientes capítulos:

En el primer capítulo se realizará un análisis del marco referencial, que incluye otros estudios referentes a la estimación de probabilidades en la cartera de los clientes, análisis del comportamiento de estos y su segmentación.

En el segundo capítulo se detallará el marco teórico compuesto por dos partes. La primera se compone de todos los fundamentos, teoremas y métodos estadísticos que sustentarán la investigación. La segunda consiste en las definiciones y conceptos que complementarán la aplicación en el ámbito profesional.

En el tercer capítulo se presentarán los resultados obtenidos del análisis y procesamiento de datos.

En el cuarto capítulo se discutirán los resultados obtenidos.

1. ANTECEDENTES

Todos los pasos y movimientos que dan las empresas dejan un rastro digital dentro de los sistemas de planificación de recursos (ERP). Si estos datos son ordenados y transformados para mostrar información relevante, es posible facilitar la toma de decisión para la organización. Esto mismo aplica con las decisiones sobre los clientes, donde las técnicas de segmentación y predicción son cada vez más acertadas.

Un estudio realizado por Aleksandrova (2018) sobre una empresa de fabricación de concreto, utilizó los datos transaccionales del ERP de la empresa para extraer variables de reciencia, frecuencia y monto de compra de sus diferentes clientes, y así calcular la probabilidad de pérdida de los clientes. Para determinar que un cliente se perdió, analizaron sus ciclos de compra, determinando que para ese mercado la inactividad de seis meses o más indica que pertenece a dicha categoría (clientes perdidos).

Basados en esta ventana temporal, dividieron los periodos de tiempo en dos. Para cada cliente calcularon la reciencia, frecuencia y monto de compra que tuvieron previo a los últimos seis meses, y determinaron por último si los clientes habían estado activos o inactivos en los últimos seis meses. De esta forma, estimar la probabilidad de pérdida de los clientes seis meses atrás, y validar este modelo con el periodo de tiempo que le procedió. Dado que la pérdida o retención de un cliente es una opción binaria, se podría considerar que la probabilidad de pérdida de un cliente es complementaria a su probabilidad de compra, lo que podría facilitar su estimación. Aun así, la

investigadora hace la recomendación de buscar otras variables que permitan la segmentación de los clientes para hacer estimaciones más precisas.

Por otro lado, en el estudio de Cuadros et al. (2017) se buscó hacer una segmentación de cartera de clientes, basada en parámetros de reciencia, frecuencia y monto de compra por medio de la aplicación de técnicas de análisis estadístico multivariado. Adicional a las variables clásicas del modelo RFM, se extrajeron los datos de utilidades, margen neto y días de crédito vencidos. En el proceso de minería de datos se procedió a normalizar los datos para comparar las variables en la misma escala de medición, para medir la covarianza, colinealidad y correlación. Se debe considerar que en este estudio no se incluyeron en el análisis algunas variables cualitativas de interés que podrían también segmentar a los clientes; sin embargo, provee un buen marco referencial para segmentar la clientela actual de la empresa.

En un estudio reciente sobre la segmentación avanzada de clientes, de Yoseph y AlMalaily (2019), se hizo el análisis de segmentación de clientela en una empresa *retail* de Malasia. El autor afirma que basarse únicamente en las variables tradicionales de segmentación, que provienen de aspectos demográficos, suele ser una fuente de errores en mercadeo, ya que no describe en su totalidad a los clientes. Por ello, también se requiere llevar a cabo un análisis de los aspectos de comportamiento de los clientes. Para el estudio de segmentación, los autores tomaron los datos de frecuencia y monto de compra para clasificar a los clientes en cuatro segmentos: frecuentes, de alto gasto, mejores clientes e inciertos. Sin embargo, por el modelo de negocio que se describe en el estudio, este estudio no tomó en consideración la variabilidad que tiene la reciencia de compra. Para tomar en cuenta una tercera variable, este tipo de segmentación resulta en una agrupación de al menos ocho estratos, que es muy numerosa para desarrollar estrategias de mercadeo.

Otro enfoque utilizado por Dogan et al. (2018) para solucionar la problemática de una óptima segmentación de clientes, basada en su comportamiento, es la aplicación de un algoritmo clasificación de K Medias. Este algoritmo hace simulaciones de clasificación para obtener las agrupaciones que disminuyan la variación total. Los autores utilizan este método con un set de datos de clientes y sus valores de RFM, con varias cantidades estratos. Finalmente, determinaron cuál disposición redujo la variación total en los datos e identificaron y qué registros pertenecen a cada estrato. Los resultados del estudio, que fue aplicado a un comercio minorista, reflejaron que con tres estratos se redujo el error total al agruparlos. La metodología descrita puede ser útil para la construcción de modelos predictivos, pues convierte tres variables cuantitativas en una sola variable categórica. Sin embargo, se debe validar si este tipo de agrupación tiene un mejor ajuste a la realidad. comparado con la regresión de las tres variables por separado.

Jain et al. (2020) efectuaron un estudio experimental en una empresa de telecomunicaciones de Estados Unidos, donde compararon diferentes modelos, dos modelos predictivos de la probabilidad de pérdida de sus clientes. Para ello, extrajeron información directamente de la plataforma de gestión de relaciones de los clientes (CRM), y la dividieron aleatoriamente en dos muestras. Una para trabajar por medio de *machine learning* y la otra, por medio de una regresión logística binaria, donde 0 es para los clientes que siguieron activos, y 1 para los que se perdieron.

El resultado de esta comparación de métodos fue similar para ambos casos; con un 85.23 % de exactitud para la regresión logística y 85.17 % de exactitud para los algoritmos de aprendizaje cerrados. Esta diferencia podría considerarse despreciable, pero ambos modelos podrían ser válidos para la predicción. Sin embargo, por la complejidad y costos de las plataformas de

inteligencia artificial, en ocasiones se recomienda obtener la información de los métodos más sencillos, como lo es la regresión logística binaria. También se debe mencionar que el modelo de negocio sobre el que se realizó este estudio (Jain et al., 2020) es manejado por contratos fijos con la clientela, con lo cual se tiene información inmediata del momento de pérdida de un cliente; en el caso de la empresa de materiales de construcción, esto es un valor fijo de referencia.

De la misma forma, en el estudio de Hargreaves (2019) se toma la pérdida o retención del cliente como una variable binaria para la construcción de un modelo binario de regresión logística. Este modelo fue elegido por ser considerado un modelo de clasificación sencillo con buenos resultados, pues en el diagnóstico el modelo obtuvo un 76.7 % de precisión. En dicho estudio se aplicó una combinación de 20 variables cualitativas y cuantitativas. Sin embargo, destacan que la principal suposición es que todas las variables independientes no tienen una multi colinealidad significativa, por lo que hace la recomendación de llevar a cabo un análisis de regresión para las variables cuantitativas, y de independencia con las variables a un nivel de significancia adecuado.

En esta línea, Senaviratna y Cooray (2019) hicieron un estudio donde buscaron diagnosticar y optimizar la multicolinealidad de un modelo de regresión logística. Los autores utilizaron el coeficiente de Pearson (r) y el factor de inflación de la varianza (VIF) para identificar las variables que están relacionadas entre sí. Destacan que, en ese momento, no había posibilidad de aumentar el tamaño de la muestra para corregir la multicolinealidad, por lo que se procedió a eliminar variables relacionadas. Para ello eliminaron la variable con mayor VIF y analizaron nuevamente la multicolinealidad; posteriormente, repitieron estos pasos hasta tener una agrupación de variables independiente.

Boateng y Abaye (2019) realizan una revisión de las recomendaciones para aplicar una regresión logística, discutiendo las mejores prácticas para la elección de variables independientes. En este, indican que el hecho de incluir una variable adicional a un modelo de regresión logística puede su exactitud al explicar cierta parte de la variación. Sin embargo, indican también que se debe ser muy cuidadoso para no incluir todas las variables que existan, pues hay posibilidades de que algunas de estas variables no tengan una correlación significativa con la variable de respuesta; esto tiende a inflar la validez aparente del modelo construido. Se recomienda ampliamente la revisión literaria de la variable a estudiar, como uno de los métodos para facilitar la elección de las variables independientes.

Un antecedente importante en la elección de las variables es el estudio de van Smeden et al. (2018); indican que el tamaño de la muestra para regresiones logísticas se debe analizar en función del número de eventos u observaciones por variable (EPV), y que hay una tendencia a asumir que si la variable tiene EPV>10 es útil para la construcción de modelos. En el análisis, los autores obtuvieron errores de calibración de sus modelos cuando las variables contaban con menos de 10 eventos por variable. También observaron una mejora de la exactitud de los modelos al aumentar el número de EPV, pero no obtuvieron una mejora de sus resultados más allá de 20 EPV.

Los estudios anteriores resaltan que, para tener un modelo de predicción con buen ajuste, deben existir una serie de variables regresoras (cualitativas o cuantitativas) que sean independientes entre sí, pero con suficiente correlación con la variable de respuesta. Ya que no hay un método exacto para elegir las variables, la mejor elección de variables es aquella que permita reducir el error al máximo posible entre los modelos desarrollados.

En un estudio sobre la pérdida de clientes en el sector de telecomunicaciones de Pakistán (Ullah et al., 2019), se construyeron varios modelos de *machine learning* para predecir la pérdida de clientes. Para determinar cuál de los modelos construidos presentaba un mejor ajuste con los clientes, se realizó una matriz de desempeño, donde compararon la exactitud, precisión, tasa de positivos y tasa de falsos positivos de cada predicción. Estas mediciones fueron obtenidas a través una matriz de confusión, donde se consideró que, si la predicción de pérdida del cliente es de 50% o más, el modelo los clasificaba como clientes a perderse, de lo contrario, serían considerados clientes a retenerse. Al comparar esta clasificación con los datos reales, midieron cuantas predicciones fueron correctas, la cantidad de falsos negativos y la cantidad de falsos positivos.

El estudio descrito aplica conceptos de informática fuera del alcance, pero se puede considerar válida la técnica de comparación de modelos. Para ello, es imperativo hacer una segmentación en la técnica de minería de datos como la descrita en el estudio de Aleksandrova. (2018)

La revisión de estudios realizados, principalmente en otros países, evidencian el alto desarrollo que existe en el área analíticas avanzadas de clientes, para desarrollar modelos predictivos en otros modelos de negocio. Para el presente estudio, serán utilizadas para la construcción de un esquema metodológico y para la construcción de un esquema de solución estadística.

2. PLANTEAMIENTO DEL PROBLEMA

2.1. Contexto general

El mantenimiento de una cartera de clientes es una de las labores de mayor impacto dentro de una organización. Sin embargo, para que esto sea efectivo, el seguimiento constante de estos clientes debe hacerse de forma estratégica, buscando predecir su comportamiento para aprovechar los momentos donde estos son más propensos a comprar, y determinar cuando están en riesgo de perderse.

El modelo de negocio a estudiar es el de una empresa minorista de productos de construcción, con base de operaciones en la ciudad de Guatemala. Por el tipo de productos que la empresa comercializa, la clientela se suele segmentar en clientes profesionales y minoristas. Los primeros son clientes que, por la naturaleza de su negocio, suelen tener una relación comercial de largo plazo con la empresa. Por lo tanto, son el segmento en el que la empresa tiene más oportunidades de construir lealtad.

Se ha identificado que hay una brecha de información sobre el comportamiento de los clientes y su frecuencia de compras, que impide hacer un análisis profundo de las estrategias de generación de lealtad. Dicha situación causa que el área estratégica de la empresa no pueda predecir los momentos críticos de los ciclos de compra de sus clientes. Por ello, todos los clientes son tratados de la misma forma sin discriminar variables útiles, como la probabilidad de compra o pérdida del cliente, su región o su segmento de mercado.

2.2. Descripción del problema

Dado que no se cuenta con estudios válidos sobre el comportamiento de los clientes profesionales, en la empresa no es posible desarrollar herramientas estadísticas que permitan impulsar técnicas predictivas sobre el comportamiento de los clientes actuales. Por este motivo, se necesita conocer cuál es la probabilidad de compra y pérdida de cada cliente profesional, tomando en consideración las variables cuantitativas y cualitativas de cada consumidor.

Por otro lado, no se ha estudiado con profundidad qué variables de segmentación de los clientes tienen una correlación significativa con la compra o pérdida de los clientes, y cuáles de estas están relacionadas entre sí. Este vacío de información existe para las variables de clasificación de los clientes, que son categóricas, como las variables cuantitativas de reciencia, frecuencia y monto de compras de cada cliente.

2.3. Formulación del problema

Al partir de la descripción del problema anterior, es posible identificar las preguntas generadoras que permiten la formulación de la investigación y su esquema de solución.

2.3.1. Pregunta central

¿Cómo se comporta la probabilidad de retención de los clientes profesionales, en función de sus variables cuantitativas y cualitativas?

2.3.2. Preguntas auxiliares

- ¿Cuáles son las agrupaciones óptimas de los clientes, en función de sus variables RFM?

- ¿Qué variables de los clientes profesionales provocan una variación significativa en la retención o pérdida de los clientes?

- ¿Cuál es el modelo óptimo para describir la probabilidad de retención o pérdida de los clientes profesionales?

2.4. Delimitación del problema

El problema abarca al segmento de clientes que están categorizados como clientes profesionales en Guatemala, que cuenten con un mínimo de dos transacciones en los últimos dos años. Todos aquellos clientes profesionales que no cumplan con este parámetro serán considerados como cartera inactiva.

3. JUSTIFICACIÓN

La presente investigación busca la construcción de un modelo de regresión logístico para determinar la probabilidad de retención de clientes profesionales en un negocio minorista. Para ello, se deben considerar múltiples variables de segmentación geográficas, demográficas y de comportamiento, que son naturalmente de tipo cualitativas. Estas variables serán extraídas de la base de datos transaccional de la empresa. En ese sentido, el presente estudio pertenece a la línea de investigación de análisis de variables categóricas.

La importancia del proyecto de investigación radica en la necesidad de la empresa de desarrollar estrategias de mercadeo directo y desarrollo de lealtad con sus clientes profesionales. Para ello, se requiere diseñar los estímulos a entregar a los clientes basado en su estatus, lo cual es posible si se conoce cuál es la probabilidad de que un cliente en particular continue activo o se pierda.

El objetivo de efectuar esta investigación es que el análisis del comportamiento de consumidores y la construcción de algoritmos de *machine learning* es un área no explotada. Por otro lado, el estudio de la correlación y regresión entre variables de tipo categórico, utilizando regresiones logísticas, no se ha desarrollado a profundidad en la empresa en cuestión.

El proceso de investigación tendrá como resultado la identificación de las variables que se encuentran correlacionadas con la retención de los clientes y la definición de un modelo que permita, dinámicamente, estimar la probabilidad de retención para la cartera activa de la empresa. De esa manera, las unidades

de negocio podrán diseñar e implementar estrategias comerciales. enfocadas en optimizar las variables que predicen la lealtad de los clientes, además de segmentar a los clientes en función de la probabilidad de qué pérdida o retención hay de estos.

Los beneficiarios de este proceso serán los equipos comerciales y de mercadeo, pues esto permitirá tener mejores resultados en las acciones estratégicas.

Los resultados del estudio son socialmente relevantes, pues servirán de referencia para el análisis del comportamiento de los clientes, en pequeñas y medianas empresas de Guatemala que no cuentan con los recursos necesarios para desarrollar técnicas de predicción como las descritas.

4. NECESIDADES A CUBRIR Y ESQUEMA DE SOLUCIÓN

La resolución de este problema estadístico le permitirá a la empresa la oportunidad de desarrollar estrategias de mercadeo directo, alineadas con la centricidad de los clientes. Esto implica ejecutar acciones de retención a los clientes con posibilidad de perderse, y buscar formas de conversión para aquellos que tienen una alta probabilidad de compra. Asimismo, permitirá desarrollar estrategias de precios diferenciados que permitan maximizar el ingreso.

Para la resolución del problema estadístico, se plantea la construcción de un modelo de regresión logístico, que permita conocer cuál es la probabilidad de que un cliente particular compre o no en determinado espacio de tiempo. Esto se hará mediante el análisis de las variables de RFM de los clientes y de sus variables de categorización, validándolo contra los resultados reales de estos clientes.

La fase inicial del proceso de extracción de información es determinar, para cada cliente profesional, si estos han tenido transacciones en los últimos seis meses. Si han tenido al menos una transacción, se marcarán como clientes retenidos; de lo contrario, serán clientes perdidos. Esta será considerada la variable de respuesta del modelo. Posteriormente, se hará un cálculo de los valores de reciencia, frecuencia y monto promedio de gasto de cada cliente de los dos años anteriores. Estas serán las variables cuantitativas que permitan medir el comportamiento de los clientes. Por último, se añadirán las variables categóricas (género, condiciones de pago, sucursales y datos demográficos,

entre otros.) que definen a estos clientes para tomarlas en cuenta en la construcción del modelo.

Para el uso de las variables cuantitativas dentro del modelo logístico, será necesario hacer un análisis de bondad de ajuste de estas variables, para poder estimar adecuadamente los parámetros de comportamiento. Asimismo, se hará un análisis de correlación entre las variables, para entender qué efecto tienen entre sí. Finalmente, los clientes se agruparán en múltiples segmentos, basado en sus variables de reciencia, frecuencia y monto promedio de gasto; para ello, se utilizará el método de simulación por un algoritmo de K Medias.

Para la construcción del modelo logístico será necesaria la definición de variables a utilizar. En este punto, se hará un análisis de independencia de cada variable categórica con la variable respuesta, incluidas los segmentos de clientes definidos en el análisis de variables de comportamiento. El resultado de este análisis permitirá conocer cuáles de estos factores contribuyen a la variación de la compra o pérdida de los clientes y, por ende, estas variables podrán ser utilizadas para la construcción del modelo logístico.

La etapa de la construcción del modelo logístico se hará con las variables definidas en el punto anterior y debe incluir una fase de validación y diagnóstico. El valor predicho de la regresión logística se convertirá en una variable dicotómica usando la siguiente regla: si la probabilidad de compra es mayor o igual al 50%, el predicho toma el valor de cliente retenido; en caso contrario, tomará el valor de cliente perdido. De esta forma, será posible analizar los predichos contra la variable de respuesta original, estimando así la precisión de la predicción del modelo.

El proceso anterior deberá repetirse de forma iterativa, pues para el modelo principal (que incluye todas las variables) será necesario evaluar la multicolinealidad mediante el factor de inflación de la varianza. Se eliminarán una a una las variables que estén añadiendo incertidumbre al modelo original, evaluando la precisión en cada paso hasta obtener el modelo con el mejor ajuste al comportamiento de la variable de respuesta.

5. MARCO TEORICO

5.1. Estadística

Salazar y Del Castillo (2018) definen la estadística como la ciencia que se encarga de recolectar, ordenar, representar, analizar e interpretar datos generados en una investigación sobre hechos, individuos o agrupaciones de estos, para realizar conclusiones o estimaciones precisas.

5.1.1. Estadística descriptiva

Los mismos autores indican que la estadística descriptiva, también conocida como matemática, es aquella que permite analizar un conjunto de datos, extrayendo conclusiones que son válidas únicamente para este conjunto de observaciones. Por ello, con los estudios realizados con este tipo de análisis, únicamente es posible describir los resultados que se obtuvieron.

5.1.2. Estadística inferencial

Por otro lado, se define a la estadística inferencial como la rama que pretende obtener estimaciones generales de una población, mediante el estudio de una muestra proveniente de la población. Para este proceso, es necesario conocer los conceptos de valores estadísticos y parámetros; inicialmente, se calculan los valores descriptivos de la muestra, obteniendo estadísticos referentes al conjunto de datos, con lo que se busca estimar los parámetros que describen a la población general.

5.1.3. Correlación entre variables

La construcción de modelos de regresión implica elegir adecuadamente las variables independientes que intervienen en el fenómeno estudiado. Para ello, se requiere conocer qué tan relacionadas están las variables independientes con la variable de estudio. La técnica de correlación se encarga de determinar cuál es el grado de asociación que existe entre dos variables, ya sea positiva o negativa, como indican Badii et al. (2014). Este análisis puede realizarse entre variables cuantitativas y cualitativas mediante diferentes métodos que se describen a continuación.

Cuando se determina la correlación entre variables, se debe prestar especial atención al tipo de variables que se examinan, pues según Hernández-Lalinde et al. (2018), la validez de las conclusiones que se extraen en una inferencia estadística depende del cumplimiento de los supuestos bajo los cuales se han construido los modelos.

5.1.3.1. Coeficiente de Pearson

También conocido como R de Pearson, es el método clásico para la medir la correlación entre variables cuantitativas. En este estudio se obtiene la magnitud, sentido y significación de la asociación de ambas variables.

5.1.3.1.1. Supuestos

Para que los resultados del análisis de correlación por medio del coeficiente de Pearson sean válidos, los datos deben cumplir con siete supuestos que se conocen por medio de un análisis previo:

- El nivel de medición de ambas variables debe ser de intervalo.
- La medición de los datos debe realizarse simultáneamente, de forma que se presentan datos pareados.
- Ambas variables deben tener un comportamiento normal.
- Debe haber ausencia de datos atípicos bivariados, pues un par de datos que se distancia considerablemente del resto de los resultados afecta seriamente el análisis de correlación.
- La asociación entre las variables a probar debe tener una tendencia lineal. También es útil realizar transformaciones matemáticas para variables que no tienen un comportamiento normal.
- Debe existir independencia entre cada par de observaciones.
- El muestreo empleado debe ser aleatorio.

5.1.3.1.2. Definición

Una vez se ha verificado el cumplimiento de los supuestos anteriormente descritos, es posible el cálculo del coeficiente de correlación de Pearson. Sea X la variable independiente e Y la variable independiente y los datos están presentados en la forma de *n* pares ordenados; según Bewick et al. (2003), el coeficiente de correlación de Pearson está dado por la siguiente ecuación:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad \text{(Ec.01)}$$

El valor obtenido de r estará siempre entre valores -1 y +1. El símbolo de este coeficiente indica el sentido de la correlación; siendo negativo para correlaciones inversas y, positivo, para correlaciones directas. Por otro lado, el valor del coeficiente indica qué tan fuerte es esta asociación; los valores más extremos indican relaciones más fuertes.

Es evidente que un valor de r cercano a 0 indica que no hay relación entre las variables; sin embargo, vale la pena repasar el supuesto de linealidad de la asociación de las variables. Dado que el modelo de Pearson para determinar la correlación está basado en un modelo lineal, un valor 0 de r indica que no existe una relación lineal, pero se debe profundizar el análisis para determinar si la relación entre las variables obedece a otro tipo de función.

Para realizar este análisis, es factible utilizar el coeficiente de Pearson con transformaciones de variables, siempre y cuando estas cumplan con los supuestos mencionados anteriormente.

5.1.3.1.3. Análisis

El valor obtenido del coeficiente de Pearson es útil para la determinación de datos muestrales; sin embargo, para hacer estimaciones de la proporción se suele utilizar una prueba de hipótesis para determinar si la asociación es significativa a no. Para ello, en su análisis de métodos de correlación, Badii et al. (2014) indica que es posible plantear las hipótesis de la siguiente forma:

$$H0: r = 0$$
$$H1: r \neq 0 \qquad \text{(Ec.02)}$$

Si la hipótesis nula es aceptada, se afirma con un nivel de significancia α que no existe correlación significativa entre las variables a nivel poblacional.

5.1.3.2. Coeficiente de Spearman

Cuando las dos variables bajo estudio no cumplen con los supuestos de normalidad y linealidad que supone el coeficiente de Pearson, una de las

alternativas no paramétricas para medir la correlación es el método de Spearman.

5.1.3.2.1. Supuestos

Para la aplicación de este método de correlación, los datos deben cumplir con los siguientes supuestos:

- Las variables deben ser de intervalo, de razón u ordinales, de forma que puedan ordenarse.
- No es necesario que se asuma una relación lineal entre las variables.
- No es requerido que las variables sigan una distribución normal.

5.1.3.2.2. Definición

El coeficiente de correlación de Spearman, denotado como $r_{spm,}$ no está basado en el valor que toma cada variable en su *i-ésima* observación, sino en la posición relativa que esta observación toma respecto al conjunto de datos. Para ello, definidas las variables x e y, se calcula el rango o posición media de cada observación; a esta variable se le denotará como R_x y R_y.

El coeficiente de Spearman estará dado por la siguiente ecuación:

$$r_{spm} = 1 - \frac{\sum_{i=1}^{n} D_i^2}{n(n^2 - 1)} \qquad \text{(Ec.03)}$$

Donde *n* es el tamaño de la muestra y, *D*, es la diferencia entre los rangos de *X* e *Y* de cada observación. Lo que algebraicamente puede definirse como:

$$D_i = R_{xi} - R_{yi} \qquad \text{(Ec.04)}$$

Al igual que el análisis del coeficiente de Pearson, el valor de r_{spm}varía desde -1 hasta +1 de forma adimensional. Siendo los valores más extremos los que describen las correlaciones entre variables más fuertes, y los cercanos a cero implican falta de correlación

5.1.3.2.3. Análisis

La correlación de Spearman también puede ser analizada por medio de pruebas de hipótesis para hacer afirmaciones correspondientes a la población. Para ello, se plantean las hipótesis de la siguiente forma:

$$H0: r_{spm} = 0$$
$$H1: r_{spm} \neq 0 \qquad \text{(Ec.05)}$$

Si la hipótesis nula es aceptada, se afirma con un nivel de significancia α que no existe correlación significativa entre las variables a nivel poblacional. La comparativa del estadístico de prueba con los valores críticos puede definirse desde la tabla de Spearman, que contiene valores estándar para un tamaño de muestra y nivel de significancia dados.

5.1.3.3. Coeficiente de Kendall

Para Serna Morales (2019), el coeficiente de Kendall, denotado τ, es una alternativa no paramétrica para el cálculo de la correlación entre variables que no presentan un comportamiento paramétrico. Se basa en los intervalos

jerarquizados de las observaciones, lo que permite que la distribución de τ sea independiente de los valores que presentan las variables x e y.

5.1.3.3.1. Supuestos

Para la aplicación del modelo de Kendall, los datos deben cumplir con los mismos supuestos utilizados para el coeficiente de Spearman, los cuales son:

- Las variables deben ser de intervalo, de razón u ordinales, de forma que puedan ordenarse.
- No es necesario que se asuma una relación lineal entre las variables.
- No es requerido que las variables sigan una distribución normal.

5.1.3.3.2. Definición

Dado que la estimación de τ de Kendall está basado en la concordancia y discordancia de sus pares ordenados. Esto lo define Serna Morales (2019) como "Sea $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ una muestra aleatoria de *n* observaciones de variables continuas o al menos ordinales, un par de observaciones (x_i, y_i) y (x_j, y_j) son concordantes si $x_i < x_j$ y $y_i < y_j$, o si $x_i > x_j$ y $y_i > y_j$". Por ende, cada par de datos será concordante o discordante, con lo que podrá calcularse el *τ* de Kendall como:

$$\tau = \frac{Nc - Nd}{Nc + Nd} \qquad \text{(Ec.06)}$$

Donde N_c denota el número de pares concordantes, y N_d denota el número de pares discordantes.

El resultado del coeficiente de Kendall será un valor entre -1 y 1, que define el grado de asociación que existe entre dos variables. Si se trata de variables independientes, el valor esperado para el coeficiente de Kendall es cero.

5.1.3.3.3. Análisis

El valor calculado de τ *también* puede ser usado como un estadístico de prueba para hacer pruebas de hipótesis que ayuden a hacer aseveraciones sobre la población. Para ello, la hipótesis a plantear sigue la siguiente lógica:

$$H0: \tau = 0$$
$$H1: \tau \neq 0 \qquad \text{(Ec.07)}$$

Si la hipótesis nula es aceptada, se afirma con un nivel de significancia α que no existe correlación significativa entre las variables a nivel poblacional. El valor crítico puede definirse por medio de la distribución normal para tamaños de muestra suficientemente grandes.

5.1.3.4. Pruebas de independencia

El análisis de correlación para variables categóricas difiere de los métodos utilizados en variables cuantitativas, ya que en estos no es posible utilizar sus valores o jerarquías para definir coeficientes de interés. El análisis de este tipo de relaciones se realiza con una tabulación ordenada de las frecuencias de los datos ordenados mediante tablas de contingencia.

Lopez-Roldan & Fachelli (2015) definen una tabla de contingencia como una tabla de frecuencias que resulta de la distribución conjunta al relacionar o

cruzar dos o más variables cualitativas. Estas pueden representar la tabulación cruzada de *n* variables; sin embargo, para el propósito de medir la independencia de los datos, se hará énfasis en las tablas de contingencia bidimensionales.

Las tablas de contingencia se completan con los totales de filas y columnas, que serán utilizados para hacer diferentes estimaciones de la distribución bivariada, observando la homogeneidad e independencia de los datos. Para ellas, es posible la implementación de pruebas de independencia como una alternativa para determinar la correlación entre variables categóricas.

Figura 1. **Tabla de contingencia bidimensional**

Estructura de Tablas de Contingencia

	Segmento			
Método de envío	Cliente	Empresa	Pequeña empresa	**Total general**
Estándar	3,165	1,864	1,144	6,173
Mismo día	272	150	93	515
Rápido	1,097	618	382	2,097
Urgente	759	417	293	1,469
Total general	5,293	3,049	1,912	10,254

Fuente: elaboración propia, hecho con Tableau.

5.1.3.4.1. Supuestos

Para aplicar una prueba de independencia es necesario que los datos cumplan con los siguientes criterios:

- Se presentan tabulados en una tabla de contingencia de dos dimensiones.
- Las variables son cualitativas (medidas a nivel nominal u ordinal, o son tratadas en esa escala de medición).

5.1.3.4.2. Definición

Las pruebas de independencia, a diferencia de los métodos de correlación para variables cuantitativas, no tiene como resultado un coeficiente que indique el grado de asociación, más bien se trata de una prueba de hipótesis basada en la distribución X^2, que permite establecer o no una relación significativa entre las variables. Para esto, se plantean la hipótesis nula e hipótesis alterna de la siguiente forma:

H0: Las variables son independientes
Ha: Las variables no son independientes, (Ec.08)
por lo que existe asociación.

Estas hipótesis serán probadas por medio de un estadístico de prueba basado en la distribución X^2, que algebraicamente está definido como:

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(f_{ij} - fe_{ij}\right)^2}{fe_{ij}} \qquad \text{(Ec.09)}$$

Donde f_{ij} es la frecuencia de la *i-ésima* fila en la *j-ésima* columna, fe_{ij} es la frecuencia de la *i-ésima* fila en la *j-ésima* columna. Esta última, puede calcularse utilizando la siguiente expresión:

$$fe_{ij} = \frac{total\ de\ fila\ i * total\ de\ columna\ j}{total\ de\ observaciones} \qquad \text{(Ec.10)}$$

Los grados de libertad que toma esta distribución bivariada están definidos por la cantidad de niveles que presentan ambas variables. Por tanto, los grados de libertad v están dados por:

$$v = (i-1) * (j-1) \qquad \text{(Ec.11)}$$

Con estos cálculos realizados, es posible concluir mediante el cálculo del p-valor, aceptando o rechazando el supuesto de independencia de datos. Por la naturaleza de este análisis, el resultado no incluye un grado de asociación con una dirección definida; únicamente es posible afirmar la dependencia o independencia con un nivel de significancia definido.

5.1.4. Regresión logística

El análisis de correlación descrito en la sección anterior es únicamente útil para conocer la relación que hay entre dos variables. Posterior a ello, es común hacer inferencias estadísticas por medio de técnicas de regresión. En este contexto, una de las tareas de mayor importancia es conocer la naturaleza de los datos que se desean explicar o predecir.

La regresión logística es un método de análisis estadístico que permite predecir, por medio de probabilidades, el resultado de una variable dependiente de tipo categórica. Aunque estos modelos son capaces de relacionar una variable dependiente politómica (que admite múltiples valores), esta es en especial potente cuando solamente existen dos alternativas (dicotómicas).

5.1.4.1. Regresión logística binomial

En su análisis, Sagaró & Zamora (2019) indican que una regresión logística expresa la probabilidad de que ocurra determinado evento en función de sus variables regresoras; y aunque es posible hacerlo con variables politómicas, es la versión dicotómica o binomial la más potente y utilizada en el ámbito de la investigación científica.

El problema de observación propuesto implica el análisis de la retención de clientes profesionales, lo cual es un escenario con dos posibles resultados: la pérdida o retención de estos. Por esta razón, la revisión documental sobre los métodos de regresión logística estará enfocada únicamente en la regresión logística binomial.

5.1.4.1.1. Elección de las variables

Una de las fases más importantes dentro del análisis de regresión es la elección de las variables regresoras que definirán el modelo. Se considera necesario realizar análisis de correlación o independencia de datos según sean los casos presentados, y utilizar dentro del análisis únicamente aquellas con una asociación significativa. Aun así, se recomienda incluir también en el proceso aquellas variables que demostraron una correlación débil contra la variable dependiente, pues, aunque en solitario tengan una débil asociación, es posible que, al ser analizadas y probadas con el resto de covariables, estas sean más importantes.

Se recomienda excluir de este análisis las variables que, por causalidad, no pueden estar relacionadas al problema de estudio, las que sean redundantes

o que estén estrechamente relacionadas, para evitar la multicolinealidad o que sean el desenlace de la variable objetivo que por probar.

5.1.4.1.2. Tratamiento de los datos

El modelo de cálculo de una regresión logística requiere acomodar la información extraída adecuadamente, según Sagaró y Zamora (2019). Para ello, los autores recomiendan las siguientes mecánicas para el tratamiento de datos previo a la construcción del modelo de regresión logística:

- El modelo está basado en el uso de variables dicotómicas que toman el valor “1” para la presencia de la característica y “0” para la ausencia de esta.
- Si las variables son nominales pero politómicas, se ve en la necesidad de convertir cada nivel de esta en una variable dummy. Por ende, si se cuenta con una variable politómica con n posibles valores, esta se convertirá en n variables dicotómicas, como se puede observar en la figura 1.
- Las variables ordinales, por otro lado, también se crean como variables dummy en cada nivel, con la diferencia que su orden si implica un orden jerárquico.
- Además, para las variables cuantitativas, se asume que cada cambio de una unidad sobre la variable regresora tiene la misma magnitud.

Figura 2. **Estructura de las variables *dummy***

Estructura de Variables *Dummy*

ID	Color	Color Azul	Color Blanco	Color Negro
1	Blanco	0	1	0
2	Azul	1	0	0
3	Negro	0	0	1
4	Azul	1	0	0
5	Negro	0	0	1
6	Blanco	0	1	0
7	Negro	0	0	1
8	Blanco	0	1	0
9	Negro	0	0	1
10	Blanco	0	1	0
11	Negro	0	0	1
12	Blanco	0	1	0

Fuente: elaboración propia, hecho con Tableau.

5.1.4.1.3. Definición

Los mismos autores indican que el modelo de regresión logística binaria, que expresa la probabilidad p de que ocurra un evento en función de ciertas variables viene dado por:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}} \qquad \text{(Ec.12)}$$

Donde p representa la probabilidad de ocurrencia del evento dicotómico representado por la variable dicotómica, $\beta_1, \beta_2, \ldots, \beta_k$ son los coeficientes de regresión asociados a cada variable $x_1, x_2, \ldots, x_k$.

La estimación de los coeficientes que se asocian a cada variable es usual realizarla mediante un método iterativo de Newton-Rhapson, utilizando un proceso de máxima verosimilitud. Para esto, se hace una transformación logarítmica dividiendo la probabilidad por su complementario de la siguiente forma:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \qquad \text{(Ec.13)}$$

Dada la cantidad de variables que se suelen utilizar para la construcción de los modelos de regresión logística, se obtiene la necesidad de implementar *software* para iterar las variables continuamente para determinar los coeficientes del modelo.

5.1.5. Validación del modelo

Una vez definido el modelo de regresión logística, es necesario desarrollar la fase de diagnóstico, donde se corrobra el cumplimiento de los supuestos, validando si otra función u otra combinación de variables describe mejor el problema planteado. Este análisis debe enfocarse en el cumplimiento de dos aspectos: primero, el modelo debe ser congruente en cuanto a las variables elegidas y su interacción, y segundo, se debe cumplir el principio de parsimonia, que busca la menor cantidad de variables para explicar el modelo.

El diagnóstico del modelo de regresión logístico debe enfocarse desde dos perspectivas. La primera es el análisis de los residuos del modelo, que pueden ser de tres tipos: estandarizados, estudentizados y de desviación; el modelo con el menor error es el que mejor describe el problema de investigación. El segundo método corresponde al análisis de las medidas de influencia que cuantifican el efecto de cada observación sobre el vector de

predicciones, como las medidas de apalancamiento del método Leverage y las medidas de distancia del método Cook.

5.2. Empresa

El desarrollo del presente trabajo de investigación se realizará en una empresa comercializadora de materiales de construcción en Guatemala.

5.2.1. Características de la empresa

El segmento con el que opera la empresa implica que hay una gran cantidad de clientes eventuales, que compran por una construcción o remodelación puntual, así como otros clientes que se dedican a prestar este tipo de servicios. La naturaleza del primer tipo de clientes no permite la gestión de la lealtad; sin embargo, para la empresa sí es un objetivo estratégico mejorar la lealtad de los clientes que están considerados como profesionales.

Para el desarrollo de lealtad, la empresa tiene planeada la generación de estrategias mercadológicas enfocadas en las necesidades de cada usuario, para lo cual requiere efectuar un análisis profundo del comportamiento de los clientes y qué tan probable es que estos se pierdan.

5.2.2. Análisis del comportamiento de compras

Según Rahim et al. (2021), el análisis del comportamiento de los clientes en negocios minoristas es una de las actividades de mayor beneficio en el ámbito del mercadeo estratégico. Indica que analizar el comportamiento de los clientes implica conocer a profundidad la forma en la que se distribuyen anuncios, espacios y la forma en que interactúan con estos elementos. Esto

último, hace que el análisis requiera de herramientas tecnológicas complejas, que normalmente no están disponibles en los comercios.

Para simplificar este análisis, es posible medir el comportamiento de compra de los clientes por medio del análisis de su comportamiento transaccional, definiendo tres variables que describen cómo un cliente se relaciona con un comercio: su frecuencia de compra, la antigüedad desde su última compra y el nivel de gasto que presenta en cada visita. A este método se le conoce como RFM.

5.2.2.1. Reciencia (R)

Este indicador se refiere al tiempo, medido en días, que ha sucedido desde la última compra del cliente. Es de especial importancia porque permite determinar qué clientes se pueden considerar como perdidos. Algebraicamente, la reciencia está definida como:

$$R = Fecha\ Act - Ult.fecha\ Compra \qquad \text{(Ec.14)}$$

El resultado de este cálculo se representa como una variable continua para facilitar su análisis estadístico.

5.2.2.2. Frecuencia (F)

La frecuencia de compra, en el análisis del comportamiento de clientes, se refiere al tiempo medio que ocurre entre cada compra. Para su cálculo, se debe definir una ventana temporal para la revisión de datos; posteriormente, se toma la diferencia entre la última y primera transacción en esa ventana temporal

y se divide por la cantidad de transacciones. Esta expresión, puede definirse como:

$$F = \frac{\acute{U}ltima\ fecha\ de\ Compra - Primera\ Fecha\ de\ Compra}{Cantidad\ de\ Transacciones} \qquad \text{(Ec.15)}$$

Para facilitar el análisis y la comparación, la frecuencia de compra debe mantener las mismas dimensionales que la variable de reciencia.

5.2.2.3. Monto promedio (M)

El monto promedio de compras es la media aritmética de los montos facturados a cada cliente. Al igual que la variable de frecuencia, se define la misma ventana temporal sobre la cual se hace la extracción de datos. Esta está definida de la siguiente forma:

$$M = \frac{Monto\ Total\ Facturado}{Cantidad\ de\ Transacciones} \qquad \text{(Ec.16)}$$

Las unidades dimensionales para esta variable deben estar en formato de moneda, para realizar estimaciones adecuadas.

5.2.2.4. Clasificación de clientes por RFM

Los resultados del análisis descrito en la sección anterior son normalmente utilizados para segmentar a los clientes por su comportamiento en tres dimensiones. Para ello, hay dos métodos comunes para hacerlo: los métodos por quintiles y la agrupación por K medias

5.2.2.4.1. Método por quintiles

Este método, según Murad (2021), se basa en la suposición de que los quintiles de reciencia, frecuencia y monto promedio tienen una diferente tasa de respuesta a los estímulos mercadológicos que se envían. Se basa en calcular los quintiles de la distribución de cada una de las variables, identificando a cada cliente desde las tres variables descritas.

Para la frecuencia, luego del cálculo de los quintiles, se asigna la clasificación 1 para los clientes que presentan un menor tiempo medio entre compras, mientras que se asigna la clasificación más baja al quintil que tiene más tiempo promedio entre transacciones. De esta forma, la clasificación de la variable frecuencia clasifica mejor a los clientes más leales.

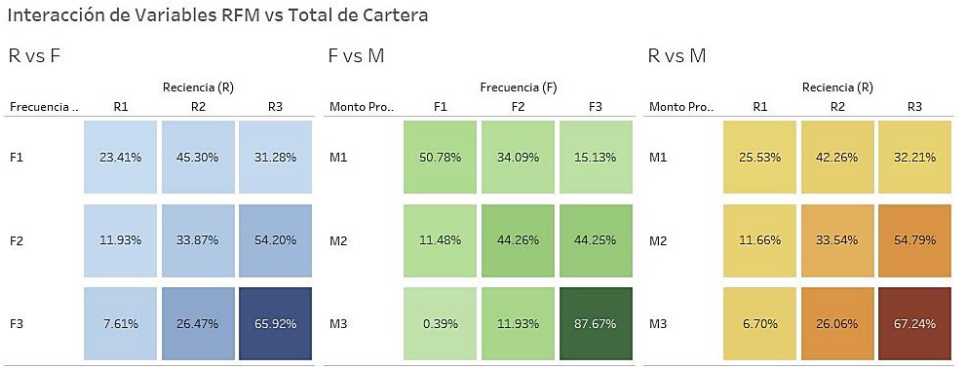
Esta misma lógica debe aplicarse para la medición de la variable de reciencia, donde se asigna la primera clasificación al quintil con menor antigüedad desde su última compra. Con ello, los clientes mejor calificados, desde esta variable, son aquellos que están más propensos a repetir su compra.

Por último, la clasificación de los clientes desde su nivel de gasto se debe hacer al inverso del resto de variables; pues se debe asignar la mejor clasificación de la variable al quintil de clientes con mayor gasto promedio.

Con esta clasificación, todos los clientes activos de la base de datos obtienen una clasificación de cada variable, y las combinaciones permiten hacer agrupaciones adecuadas de los clientes. Por ejemplo, los clientes clasificados como R1, F1 y M1 son aquellos de mayor interés para la empresa. Como se

puede observar en la siguiente figura, donde se analiza la interacción entre cada par de variables.

Figura 3. **Análisis de interacción de variables RFM**



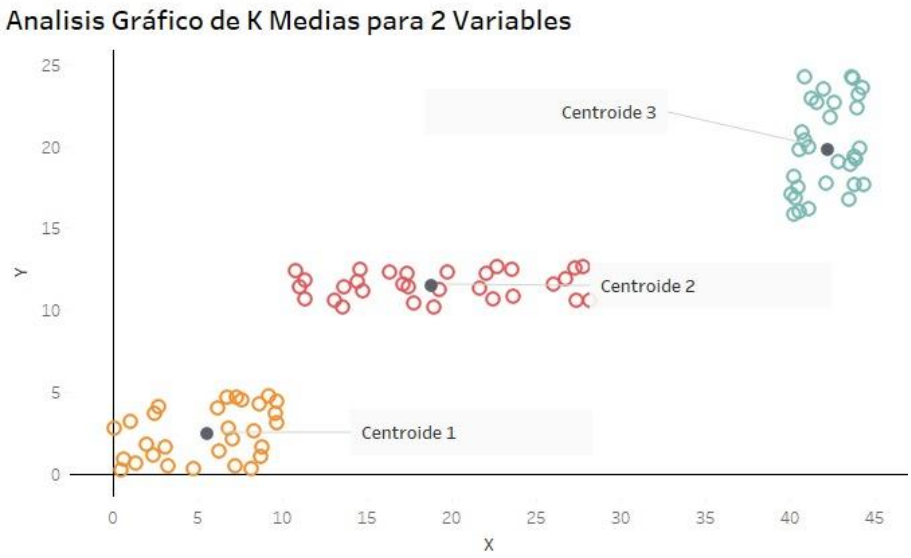
Fuente: elaboración propia, hecho con Tableau.

5.2.2.4.2. Método por *K* medias

El método de segmentación por K medias es un algoritmo no supervisado para el análisis de series extensas de datos, como indican Anitha y Patil (2022). Este algoritmo identifica K centroides dentro de los datos y asigna cada registro de los datos al centroide más cercano. Si se hace un análisis de esto en dos dimensiones, se podría observar que las agrupaciones se asemejan a la figura 2, donde cada color representa un segmento asociado a su centroide.

Dado que los centroides dependen de los puntos que fueron asociados a este, el algoritmo debe hacer múltiples iteraciones hasta encontrar la agrupación de datos que disminuya el error total. Esta iteración se puede traducir en la agrupación de los clientes con mayor similitud.

Figura 4. **Análisis gráfico de K medias para dos variables**



Fuente: elaboración propia, hecho con Tableau.

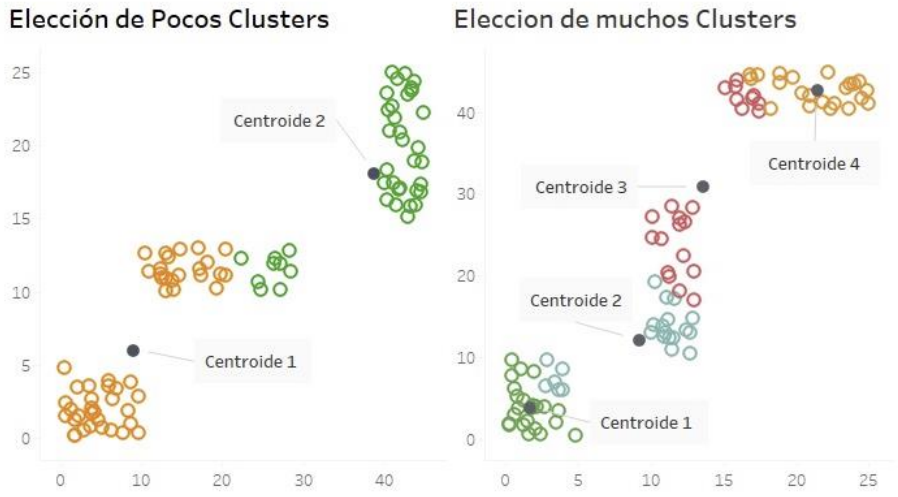
Para realizar la clasificación de los clientes, el algoritmo presenta variables o dimensiones, y busca construir k grupos donde se minimiza la suma de distancias de los objetos a su centroide. Esta distancia se puede calcular mediante la distancia euclidiana, que se define matemáticamente de la siguiente forma:

$$s = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \qquad \text{(Ec.17)}$$

Donde q_i es el valor de la variable y p_i es el valor del centroide. Esto es calculado para todas las variables de segmentación.

La cantidad de segmentos K que se utiliza para segmentar debe ser una definición del investigador; sin embargo, cabe mencionar que a medida que aumenta la cantidad de segmentos, naturalmente disminuye el error total de segmentación artificialmente. Según Hernández et al. (2019), esta es una de las principales desventajas de esta técnica de clasificación.

Figura 5. **Elección incorrecta de clústeres**



Fuente: elaboración propia, hecho con Tableau.

Elegir la cantidad adecuada de segmentos sobre los cuales optimizar, es una de las tareas críticas en el contexto del algoritmo de k medias. Esto es particularmente sencillo si se trata de un set de datos con dos o tres dimensiones, pues el análisis se puede realizar de forma visual. Sin embargo, si incrementan las variables, ya no es posible representar este modelo en 3 dimensiones, por lo que se debe recurrir a otros métodos de amplia aceptación: Método Silhouette y Método Elbow.

5.2.2.4.3. Método Silhouette

Este método utiliza un coeficiente conocido como Silhouette, que está definido como la diferencia entre la distancia promedio de los elementos a su centroide más cercano y la distancia intra clúster dividido por el máximo de los dos. Se itera este algoritmo con diferentes valores de K; cuando el coeficiente de Silhouette se maximiza, se obtiene la cantidad óptima de centroides sobre los cuales maximizar

$$CS = \frac{1}{N}\sum_{i=1}^{N}\frac{b(x) - a(x)}{\max\{a(x), b(x)\}} \qquad \text{(Ec.18)}$$

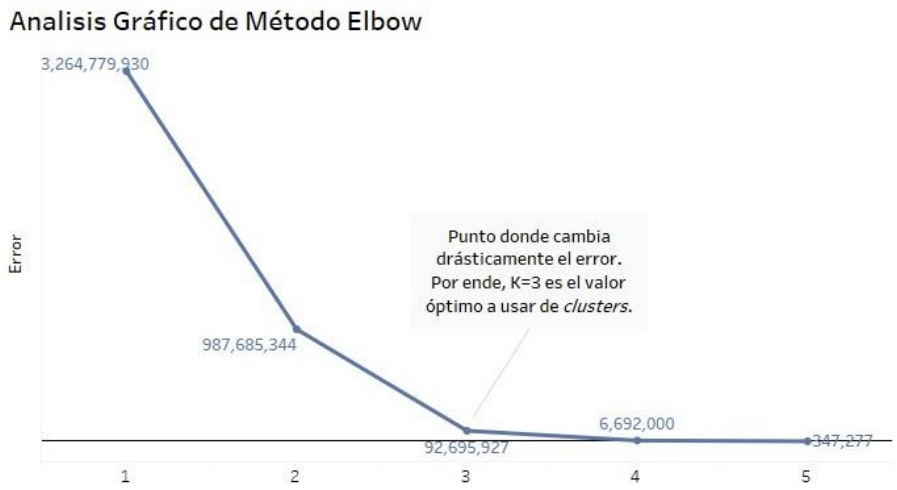
Donde *CS* denota el valor del coeficiente de Silhouette, $b(x)$ es la distancia media de los centroides y $a(x)$ es la distancia intra clúster y *N* denota cada grupo sobre los cuales se está haciendo el cálculo del coeficiente.

5.2.2.4.4. Método Elbow

Este método calcula la suma de cuadrados del error como la métrica principal de ajuste para esta agrupación. Naturalmente, al incrementar la cantidad de segmentos o clústeres la distancia entre los puntos y sus correspondientes centroides disminuye, por lo que la suma de los cuadrados del error también disminuye.

A medida que se itera para los diferentes valores de K, el error total del modelo se reduce drásticamente para los valores más pequeños, pero a medida que el valor de K incrementa, la reducción del error se hace más pequeña. En un análisis gráfico de este comportamiento, se obtiene un punto de inflexión en el gráfico, que indica el valor óptimo de centroides a utilizar.

Figura 6. **Análisis gráfico con el método Elbow**



Fuente: elaboración propia, hecho con Tableau.

5.2.3. Segmentación de la clientela

Para Gajanova et al. (2019), identificar que tan leales a la marca son los clientes es una tarea de especial interés para el éxito mercadológico y el tratamiento adecuado de estos. Por ende, la tarea de segmentación de los clientes basado en la actividad de estos es de gran valor. Este tipo de segmentación se logra mediante el análisis de variables de comportamiento descritos en el método RFM.

Sin embargo, los autores indican que es un error común tratar de analizar el comportamiento de los clientes únicamente por la forma en la que estos interactúan transaccionalmente. En realidad, se requiere mucho más contexto referente a otras variables, que son útiles para clasificar a los clientes. Uno de los métodos que los autores toman para resolver este problema, es

analizando a los clientes desde cuatro tipos de segmentación que se describen a continuación.

5.2.3.1. Segmentación geográfica

La segmentación de los clientes por su ubicación geográfica se refiere, principalmente, a agrupar a los clientes por las regiones en las que estos se encuentran. La lógica detrás de este tipo de segmentación es que hay barreras físicas que pueden diferenciar el comportamiento de los clientes. Por ejemplo, el tamaño del mercado o los incentivos fiscales pueden ser determinantes en el comportamiento de los clientes.

Entre las variables más comunes de segmentación geográfica se encuentran: país, ciudad, idioma, clima, densidad y población, entre otras. De estar disponibles todas estas variables de segmentación, son útiles para el análisis. Sin embargo, debe considerarse un análisis de independencia de datos entre la variable de segmentación y la variable objetivo, para establecer si estas representan un factor determinante en el comportamiento de los clientes.

5.2.3.2. Segmentación demográfica

Los criterios demográficos de segmentación suelen ser los más utilizados para segmentar a los clientes. En el contexto de comercios minoristas, se busca encontrar el género, edad, estilo de vida o nivel socioeconómico de los clientes. Este enfoque funciona si se estudia a clientes individuales, pero el enfoque de clientes profesionales del presente estudio implica que la mayoría de los clientes por analizar son empresas, por lo que estos factores no son útiles para la segmentación.

Esto obliga a tomar en consideración otros factores de segmentación, como los descritos por Möllering en (2018). Se sugiere que para la segmentación se tome en cuenta principalmente las características de la organización, como el tamaño de la organización, segmento de mercado en el que opera o incluso la tecnología con la que operan.

En segundo plano, se debe tomar en consideración las características demográficas de los compradores con quienes la empresa tiene contacto; para ello se requiere tener un sistema de manejo de relaciones con clientes (CRM) con esta información.

Por último, también se recomienda evaluar las condiciones financieras del cliente, tales como el estatus de cobro, límites de crédito y términos de pago.

5.2.3.3. Segmentación psicográfica

Este tipo de segmentación tiene como objetivo explicar las diferencias en las que actúan los clientes, basados en sus predisposiciones sociales y psicológicas, y trata de agrupar a clientes con características demográficas y geográficas similares que se comportan de diferente manera. Entre este tipo de variables, se puede explorar todo lo que conlleve el comportamiento de los clientes en torno a sus transacciones.

Mucha de esta información es difícil de conseguir para toda la población de clientes; sin embargo, hay otras variables de segmentación que pueden obtenerse desde el análisis de los resultados transaccionales. En el análisis por estas variables puede considerarse la modalidad de entrega de los productos, la amplitud de categorías que compran o la forma de pago, entre otras.

5.2.3.4. Segmentación por satisfacción y lealtad

Gajanova et al. (2019) sugiere que uno de los principales predictores de la lealtad de los consumidores es la satisfacción de estos. Es por ello que propone segmentar a los clientes en cuatro diferentes categorías, según su satisfacción y percepción hacia la marca:

- Defensores: se trata de los clientes que son leales a la marca y se encuentran muy satisfechos con los servicios que se prestan.

- Rehenes: son clientes que no están satisfechos con los servicios que presta la marca, pero son leales por algún factor externo.

- Mercenarios: se clasifican de esta forma a los clientes que están satisfechos, pero no son leales a la marca, por lo que están propensos a perderse.

- Terroristas: son clientes que no son leales ni están satisfechos con los servicios.

Contar con este tipo de segmentación es útil para predecir los movimientos futuros de los clientes. Sin embargo, se requiere un profundo entendimiento de los clientes, ya que el nivel de satisfacción de los clientes se puede conocer únicamente a través de programas de encuestas con los clientes.

6. PROPUESTA DE ÍNDICE DE CONTENIDOS

7. METODOLOGÍA

7.1. Características del estudio

El enfoque del estudio será cuantitativo, debido a que se tomarán mediciones de variables continuas por medio del método RFM, y estas se combinarán con otros factores categóricos, como la dirección, método de pago y otras variables de clasificación de los clientes para estimar la probabilidad de retención de estos.

El alcance del proyecto será descriptivo, pues el presente trabajo de investigación implica la construcción de un modelo estadístico funcional, para predecir la retención de clientes profesionales. Además, será de tipo correlacional porque se estudiará la relación que tienen las variables del estudio con la variable de respuesta.

El diseño adoptado será no experimental (observacional), ya que los datos para el estudio serán obtenidos por medio de minería de bases de datos. La información del comportamiento de los clientes se analizará en su estado original sin ninguna manipulación; además, será transversal, pues se hará un análisis de los datos históricos transaccionales de los clientes en un punto en específico, sin evaluar la evolución de estos.

7.2. Unidades de análisis

La población en estudio estará constituida por los clientes profesionales de la empresa minorista de materiales de construcción, de la cual se extraerá la

totalidad de la población de la base de datos transaccional para su estudio y estimación.

7.3. Variables

A continuación, se muestra el detalle de la operativización de las variables, con su correspondiente definición a nivel operativo.

Tabla I. **Operativización de variables**

Variable	**Definición teórica**	**Definición operativa**
Estado del cliente	Es la variable de respuesta del modelo. Define si el cliente se perdió o se retuvo en un periodo determinado. De tipo cuantitativa dicotómica.	Si el cliente ha tenido al menos una compra desde la fecha de referencia se considerará como cliente activo. Por el contrario, será considerado perdido.
Reciencia (R)	Es la cantidad de días transcurridos desde la última transacción del cliente. Variable de tipo cuantitativa de razón.	Diferencia entre la fecha de cálculo y la última fecha de transacción, medido en días.
Frecuencia (F)	Es la cantidad de días promedio que hay entre cada transacción. Variable de tipo cualitativa de razón.	Se definirá como el recuento de las facturas, medido en días

Continuación de la tabla I.

Variable	**Definición teórica**	**Definición operativa**
Monto Promedio (M)	Es la cantidad promedio que gasta un cliente en cada visita. Variable de tipo cualitativa de razón.	Se calculará como el promedio del total de las facturas emitidas al cliente, será medido en la divisa local.
Departamento	Es el departamento en el cual operan los clientes. Variable categórica de tipo nominal	Se obtendrá por medio de la dirección fiscal registrada por el cliente.
Condición de Pago	Es la forma en la que el cliente paga por los productos obtenidos en cada transacción. Variable de tipo categórica con escala nominal.	Se extraerá de la base de datos el tipo de pago del cliente.
Forma de entrega	Es la forma en la que el cliente recibe su producto, pudiendo tener 3 diferentes valores: ruta, recolección en tienda y múltiples. Variables de tipo cualitativo nominal.	Se extraerá la forma de entrega predeterminada de cada cliente por medio de la minería de datos

Continuación de la tabla I.

Variable	Definición teórica	Definición operativa
Cliente Multicategoría	Mide si el cliente está asociado a la empresa transaccionando solo un tipo de productos o múltiples categorías de estos. Esta variable estará definida como dicotómica.	Se hará un conteo de las categorías compradas por el cliente en la ventana de tiempo. SI este es mayor que uno, se considerará como cliente multicategórico.

Fuente: elaboración propia, hecho con Microsoft Word.

7.4. Fases del estudio

La construcción del presente estudio estará compuesta por las etapas descritas a continuación:

7.4.1. Fase uno: revisión de literatura

En esta fase, se hará una recopilación bibliográfica que será de utilidad para fundamentar los temas que se analizarán en esta investigación.

Esta revisión estará centrada en los conocimientos estadísticos, como el análisis de correlación y construcción de modelos de regresión logísticos, y una sección enfocada en la temática específica de la empresa y en el análisis del comportamiento de los clientes.

7.4.2. Fase dos: minería y limpieza de datos

Para la segunda fase del proceso de investigación, la empresa proporcionará acceso a sus bases de datos transaccionales. Con ello, se realizarán tareas de minería de datos con Microsoft SQL Server, para extraer las variables de segmentación y de comportamiento de los clientes, utilizando documentos transaccionales del 2021 al 2022.

El estudio utilizará la observación directa como método principal, para la recolección de información desde un punto de vista no participante, ya que estará basado íntegramente en el análisis de los datos transaccionales de la empresa

Para la extracción de los datos, se utilizará un modelo de entrenamiento, utilizando las transacciones de los últimos seis meses para medir el comportamiento de la variable objetivo (retención o pérdida del cliente), y las transacciones anteriores para obtener las variables regresoras.

7.4.3. Fase tres: análisis de correlación

Para la construcción de modelos adecuados, se debe conocer a detalle qué interacción tienen las variables regresoras. Para ello, en esta fase, se tomarán los datos extraídos mediante la minería de datos, para analizar la correlación que hay entre las variables.

Inicialmente, se calculará el coeficiente de correlación entre las variables cuantitativas, como las obtenidas mediante el método RFM, para analizar el grado en que estas interactúan entre sí. Este análisis permitirá eliminar variables que no aporten información relevante al modelo.

Las variables cuantitativas resultantes del análisis anterior serán agrupadas mediante el algoritmo de *K* medias, lo que permitirá la generación de una variable cualitativa que permita el análisis de independencia con la variable objetivo.

Por último, se aplicarán pruebas de independencia a las variables regresoras con la variable dependiente. Esto permitirá analizar el grado de asociación que existe entre ellas y elegir las variables adecuadas en el proceso de construcción.

7.4.4. Fase cuatro: construcción del modelo de regresión

Esta fase tomará las variables que durante el análisis de correlación probaron ser significantes en su efecto sobre la variable de respuesta, con las que se construirá un modelo de regresión logística binomial utilizando *software* R. Adicionalmente, se realizará un análisis de multicolinealidad para determinar cuáles de las variables elegidas aportan poca información al modelo de datos.

Con este análisis de multicolinealidad se harán múltiples iteraciones, eliminando una a una las variables que aporten menos información al modelo. Con ello, se obtendrá una serie de modelos con diferentes combinaciones de variables, que serán comparados entre sí con criterios de información. Por último, se elegirán los tres modelos con mejor ajuste para realizar el análisis del error de pronóstico para estos modelos.

7.4.5. Fase cinco: análisis de resultados

La quinta fase del estudio implica analizar los resultados obtenidos en cada modelo de regresión mediante el análisis de confusión de cada modelo.

Para ello, se aplicarán los modelos elegidos en la fase anterior a los datos del modelo de entrenamiento, para obtener un valor resultante de la probabilidad de retención de los clientes. Si la probabilidad de retención calculada por el modelo es mayor a 50 %, se considerará que el modelo predice la retención de este.

Los resultados calculados serán comparados con los resultados reales del modelo de entrenamiento mediante una matriz de confusión, denotando la cantidad de valores correctamente predichos y aquellos equivocados. Con ello, se analizará la exactitud de predicción de los modelos para elegir el modelo que tenga menor error de predicción.

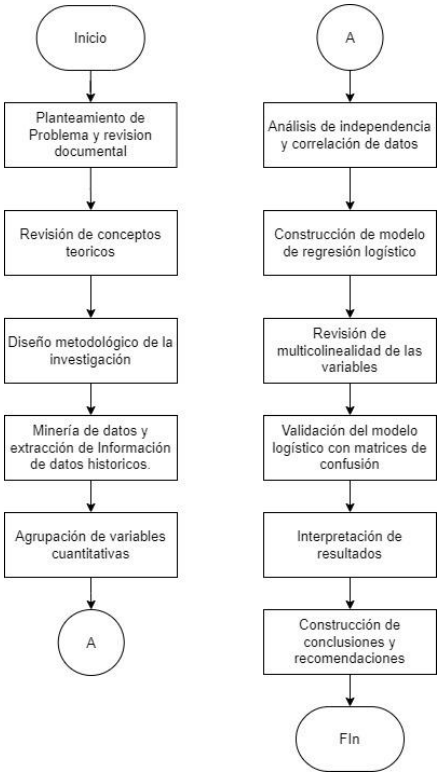
7.4.6. Fase seis: redacción de informe final y presentación de resultados

La última fase del estudio consiste en presentar los resultados obtenidos de la investigación, así como la redacción de conclusiones, recomendaciones e informe final.

7.5. Flujograma del proceso de investigación

A continuación, se detalla el flujograma que describe las fases del proceso de desarrollo e investigación.

Figura 7. **Flujograma del proceso de investigación**



Fuente: elaboración propia hecho con Diagrams.net.

8. TÉCNICAS DE ANÁLISIS DE INFORMACIÓN

A continuación, se detallan las técnicas a utilizar para llevar a cabo el estudio de investigación:

8.1. Minería y extracción de datos

La extracción de datos se realizará desde la base de datos transaccional de la empresa. La primera fase será extraer las variables cualitativas de cada cliente, mientras que la segunda tiene como objetivo extraer las variables de comportamiento de los clientes profesionales. Para ello, se tomará como referencia una fecha de seis meses previo a la fecha de cálculo; las variables del método RFM serán calculadas con las transacciones ocurridas en el año anterior a esta referencia. Por otro lado, la variable de respuesta se obtendrá analizando el comportamiento de los clientes en los seis meses posteriores a esta referencia.

8.2. Algoritmo de agrupación por K medias

Se utilizará el algoritmo de simulación para hacer una agrupación óptima de los clientes, basado en sus variables RFM, y poder utilizar estas agrupaciones o clústeres dentro del modelo de regresión logística.

8.3. Pruebas de independencia

Las variables extraídas durante la fase de minería de datos serán representadas en tablas de contingencia con la variable de respuesta, para

realizar pruebas de independencia de datos y determinar si hay un efecto significativo de las variables independientes.

8.4. Regresión logística

Se utilizará *software* R para la construcción de los modelos de regresión logística binomiales, utilizando las variables que son consideradas como significativamente relacionadas a la variable de respuesta y analizar también la multicolinealidad de las variables utilizadas.

8.5. Evaluación de modelos

Por último, se hará una evaluación del modelo obtenido en la técnica de regresión logística. Esto se hará con una matriz de confusión, donde se hará una tabulación cruzada de la variable predicha con los valores reales de la variable, determinando así la cantidad de falsos positivos y falsos negativos que entregó el modelo.

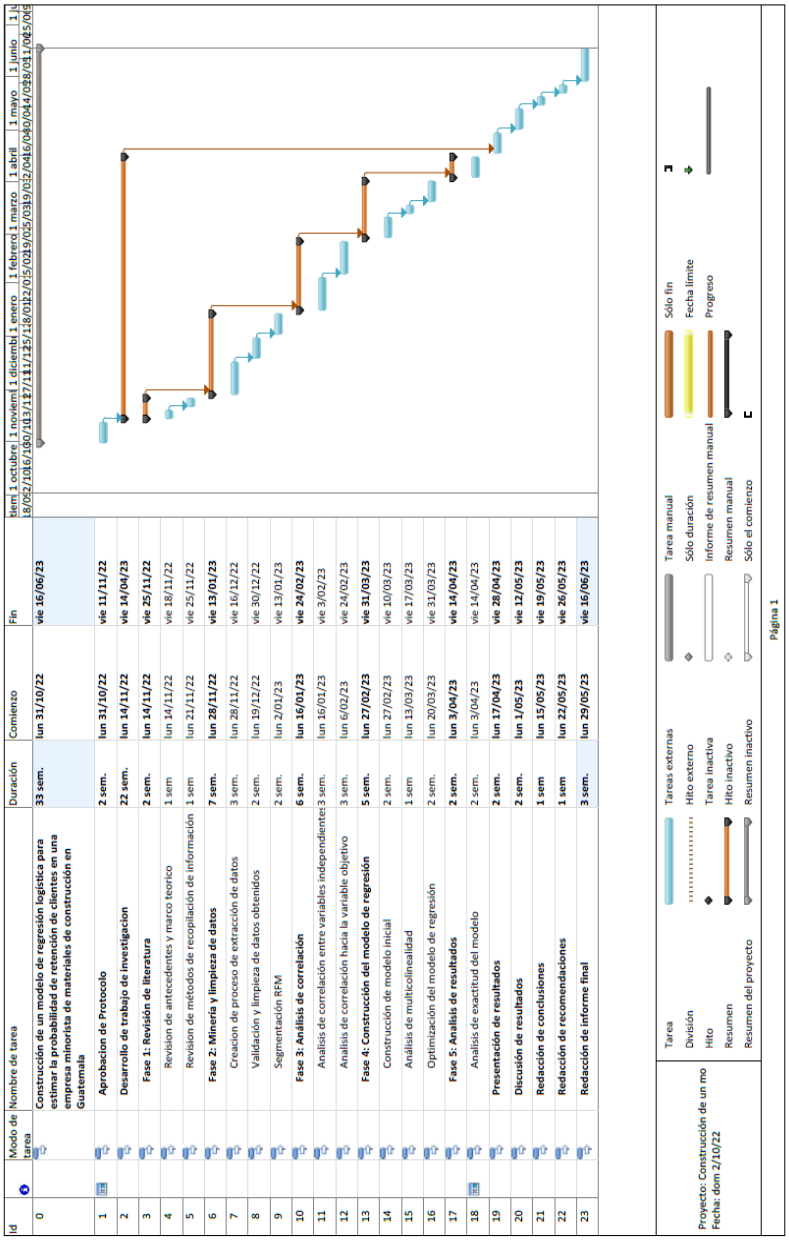
8.6. Software

El presente estudio utilizará tres herramientas de *software* para desarrollar el esquema de solución propuesta.

La minería de datos será ejecutada por medio de consultas de Microsoft SQL Server 2012; el modelado de los datos se realizará en Rstudio, y la presentación del modelo se realizará en Tableau.

9. CRONOGRAMA

Figura 8. **Cronograma**



Fuente: elaboración propia, hecho con Microsoft Project.

10. FACTIBILIDAD DEL ESTUDIO

10.1. Recurso humano

Se refiere a las horas hombre que serán invertidas en el desarrollo de la investigación. Entre estos recursos se puede destacar el tiempo del investigador. Por la duración del proyecto, se estiman 300 horas de tiempo del investigador.

10.2. Recursos financieros

A continuación, se muestra el resumen del presupuesto asignado al proyecto:

Tabla II. **Presupuesto asignado al proyecto de investigación**

Tipo de Recurso	**Descripción**	**Cantidad**	**Monto Total**
Humano	Horas invertidas en la investigación	300	Q15,000.00
Tecnológico	Licencias de *software* necesarias para la extracción de datos	3	Q 240.00
Tecnológico	Acceso a Internet de alta velocidad	6	Q 600.00

Continuación de la tabla II.

Tipo de Recurso	Descripción	Cantidad	Monto Total
Materiales	Gastos por impresión de documentos	1	Q 800.00
Transporte	Gastos por transporte y depreciación de vehículo	6	Q 1,200.00
Otros	Imprevistos	1	Q 500.00
		Total	**Q18,550.00**

Fuente: elaboración propia, hecho con Microsoft Word.

Los elementos presentados en la tabla anterior serán cubiertos íntegramente por el estudiante investigador.

10.3. Recursos tecnológicos

Se refiere a las herramientas de tecnología que serán necesarias para la ejecución de la presente investigación. Entre estas, se menciona el licenciamiento de Microsoft SQL, Software R, así como una red de internet con suficiente capacidad de respaldo.

10.4. Acceso a información y permisos

Para la obtención de datos será necesario el acceso a la base de datos de la empresa minorista de materiales de construcción, con un acceso libre a la información de sus tablas. Para ello, se solicitará un acceso a la base de datos transaccional de la empresa, firmando un acuerdo de buen uso de los datos, así como un acuerdo de confidencialidad para no revelar datos importantes de operación.

10.5. Equipo e infraestructura

Se refiere al equipo de cómputo que será empleado para el presente trabajo de investigación. Los requerimientos mínimos del equipo implican tener un procesador Core i5, con suficiente espacio en su memoria para la ejecución de algoritmos de simulación de datos.

Asimismo, será necesaria la instalación de una red privada virtual (VPN) al equipo remoto, para poder ejecutar las tareas de recolección de datos, así como el cálculo del modelo desde R.

REFERENCIAS

1. Aleksandrova, Y. (julio, 2018). APPLICATION OF MACHINE LEARNING FOR CHURN PREDICTION BASED ON TRANSACTIONAL DATA (RFM ANALYSIS). *SGEM International Multidisciplinary Scientific GeoConference EXPO Proceedings*. Congreso llevado a cabo en Sofia, Bulgaria. Recuperado de https://doi.org/10.5593/sgem2018/2.1/s07.016

2. Anitha, P. & Patil, M. M. (mayo, 2022). RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University - Computer and Information Sciences*, *34*(5), 1785-1792. Recuperado de https://doi.org/10.1016/j.jksuci.2019.12.011

3. Badii, M. H., Guillen, A., Lugo, O. P. & Aguilar, J. J. (agosto, 2014). Correlación No-Paramétrica y su Aplicación en la Investigaciones Científica. *International Journal of Good Conscience*., 31-40. http://www.spentamexico.org/v9-n2/A5.9%282%2931-40.pdf

4. Bewick, V., Cheek, L. & Ball, J. (noviembre, 2003). Statistics review 7: Correlation and regression. *Critical Care*, *7*(6), 451. https://doi.org/10.1186/cc2401

5. Boateng, E. Y., & Abaye, D. A. (noviembre, 2019). A Review of the Logistic Regression Model with Emphasis on Medical Research.

Journal of Data Analysis and Information Processing, *07*(04), 190–207. https://doi.org/10.4236/jdaip.2019.74012

6. Cuadros López, L. J., Gonzales Caicedo, C., & Jiménez Oviedo, P. C. (octubre, 2017). Análisis multivariado para segmentación de clientes basada en RFM. *Tecnura, 21*(54), 41–51. Recuperado de https://doi.org/10.14483/22487638.12957

7. Dogan, O., Aycin, E., & Bulut, Z. (2018). CUSTOMER SEGMENTATION BY USING RFM MODEL AND CLUSTERING METHODS: A CASE STUDY IN RETAIL INDUSTRY. *International Journal of Contemporary Economics and Administrative Sciences, 8*(1), 1–19. Recuperado de https://doi.org/10.5930/issn.1925-4423

8. Gajanova, L., Nadanyiova, M. & Moravcikova, D. (2019, 1 marzo). The Use of Demographic and Psychographic Segmentation to Creating Marketing Strategy of Brand Loyalty. *Scientific Annals of Economics and Business, 66*(1), 65-84. Recuperado de https://doi.org/10.2478/saeb-2019-0005

9. Hargreaves, C. A. (diciembre, 2019). A Machine Learning Algorithm for Churn Reduction & Revenue Maximization: An Application in the Telecommunication Industry. *International Journal of Future Computer and Communication, 8*(4), 109–113. Recuperado de https://doi.org/10.18178/ijfcc.2019.8.4.550

10. Hernandez, J., Tello, E., Marin, H. & Romero, G. (2019). APLICACIÓN DE TÉCNICAS DE APRENDIZAJE NO SUPERVISADO PARA LA AGRUPACIÓN DE TRAZAS EN EL DOMINIO DE MINERÍA DE PROCESOS. *Pistas Educativas, 41*(133), 356-374. Recuperado de http://pistaseducativas.celaya.tecnm.mx/index.php/pistas

11. Hernández-Lalinde, J., Espinosa-Castro, J., Penaloza-Tarazona, M., DíazCamargo, E., Bautista-Sandova, M., Riaño-Garzón, M. & Chacón Lizarazo, O. (2018). Sobre el uso adecuado del coeficiente de correlación de Pearson: verificación de supuestos mediante un ejemplo aplicado a las ciencias de la salud. *Archivos Venezolanos de Farmacología y Terapéutica, 37*(5), 552-570. Recuperado de https://www.redalyc.org/journal/559/55963207020/55963207020.pdf

12. Jain, H., Khunteta, A., & Srivastava, S. (2020). Churn Prediction in Telecommunication using Logistic Regression and Logit Boost. *Procedia Computer Science, 167*, 101–112. Recuperado de https://doi.org/10.1016/j.procs.2020.03.187

13. Lopez-Roldan, P. & Fachelli, S. (2016). *METODOLOGÍA DE LA INVESTIGACIÓN SOCIAL CUANTITATIVA*. Barcelona, España: Universitat Autònoma de Barcelona. Recuperado de https://ddd.uab.cat/pub/caplli/2015/131469/metinvsoccuan_cap3-6a2015.pdf

14. Mendivelso, F. & Rodríguez, M. (junio, 2018). Prueba Chi-Cuadrado de independencia aplicada a tablas 2xN. *Revista Médica Sanitas, 21*(2), 92-95. Recuperado de https://doi.org/10.26852/01234250.6

15. Möllering, L. (2018). *A customer segmentation sequence for B2B markets based on levels of market orientation of firms* (tesis de maestría). University of Twente, Países Bajos.

16. Murad, H. (2021). *Marketing Automation Customers Segmentation* (tesis de maestría). Rochester Institute of Technology, Estados Unidos.

17. Rahim, M. A., Mushafiq, M., Khan, S. & Arain, Z. A. (julio, 2021). RFM-based repurchase behavior for customer classification and segmentation. *Journal of Retailing and Consumer Services, 61*, 102566. Recuperado de https://doi.org/10.1016/j.jretconser.2021.102566

18. Sagaró, N. & Zamora, L. (2019). Análisis estadístico implicativo versus Regresión logística binaria para el estudio de la causalidad en salud. *Multimed, 23*(6), Recuperado de http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1028-48182019000601416

19. Del Castillo Galarza, R. y Salazar Pinto, R. (2018). *Fundamentos básicos de estadística*. Quito, Ecuador: Del Castillo Galarza, Raúl Santiago

20. Serna Morales, J. K. (2019). *Comparación de algunas estimaciones del τ de Kendall para datos bivariados con censura a intervalo* (tesis de maestría). Universidad Nacional de Colombia, Colombia.

21. Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U., & Kim, S. W. (2019). A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector. *IEEE Access, 7*, 60134–60149. Recuperado de https://doi.org/10.1109/access.2019.2914999

22. Van Smeden, M., Moons, K. G., de Groot, J. A., Collins, G. S., Altman, D. G., Eijkemans, M. J., & Reitsma, J. B. (2018). Sample size for binary logistic prediction models: Beyond events per variable criteria. *Statistical Methods in Medical Research, 28*(8), 2455–2474. Recuperado de https://doi.org/10.1177/0962280218784726

23. Yoseph, F., & AlMalaily, M. (2019). NEW MARKET SEGMENTATION METHODS USING ENHANCED (RFM), CLV, MODIFIED REGRESSION AND CLUSTERING METHODS. *International Journal of Computer Science and Information Technology, 11*(01), 43–60. Recuperado de https://doi.org/10.5121/ijcsit.2019.11104

APÉNDICE

Apendice 1. **Matriz de Coherencia**

Problema de Investigación	**Preguntas de Investigación**	**Objetivos**	**Procedimiento y Técnicas**
No hay información confiable sobre el comportamiento de los clientes profesionales	¿Cómo se comporta la probabilidad de retención de los clientes profesionales en función de sus variables cuantitativas y cualitativas?	Construir un modelo de regresión logística para estimar la probabilidad de retención de clientes profesionales en una empresa minorista de materiales de construcción en Guatemala.	Extracción de datos de clientes para el análisis del comportamiento desde sus variables de RFM.
No se conoce cuáles son los grupos óptimos las variables de RFM en clientes profesionales.	¿Cuáles son las agrupaciones óptimas de los clientes en función de sus variables RFM?	Agrupar a los clientes profesionales en segmentos similares basado en las variables de reciencia, frecuencia y monto de compras aplicando métodos de simulación por K Medias.	Aplicar algoritmo de simulación por K Medias
Se desconoce qué segmentos de mercado tienen una frecuencia de compra diferente	¿Qué variables de los clientes profesionales provocan una variación significativa en la retención o pérdida de los clientes?	Identificar las variables de clientes que sí infieren en la pérdida o retención de clientes usando pruebas de independencia y pruebas de correlación.	Pruebas de correlación e independencia de datos.

Continuación del apéndice 1.

Problema de Investigación	**Preguntas de Investigación**	**Objetivos**	**Procedimiento y Técnicas**
No se conoce la probabilidad de compra de los clientes profesionales	¿Cuál es el modelo óptimo para describir la probabilidad de retención o pérdida de los clientes profesionales?	Relacionar las variables cuantitativas y cualitativas de los clientes profesionales construyendo un modelo de regresión logística que permita cuantificar las probabilidades de pérdida y compra de cada cliente profesional	Aplicación de regresión logística bivariada, criterios de información de modelos, matrices de confusión.

Fuente: elaboración propia, hecho con Microsoft Word.